DOCUMENT RESUME

ED 431 032 TM 029 866

AUTHOR Sykes, Robert C.; Heidorn, Mark; Lee, Guemin
TITLE The Assignment of Raters to Items: Controlling for Rater Effects.
PUB DATE 1999-04-00
NOTE 37p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Constructed Response; Elementary School Students; Elementary Secondary Education; *Evaluators; High School Students; Mathematics Tests; Reading Tests; *Scores; State Programs; *Test Items; Testing Programs
IDENTIFIERS *Rater Effects

ABSTRACT

A study was conducted to evaluate the effect of different modes (modalities) of assigning raters to test items. The impact on total constructed response (c.r.) score, and subsequently on total test score, of assigning a single versus multiple raters to an examination reading of a student's set of c.r. responses was evaluated for several mixed-item format tests. Samples of approximately 2,000 students were obtained from a state mathematics field test at each of grades 5, 8, and 10 and from a reading field test at each of grades 4, 8, and 10. Item responses for c.r. items for each selected student in the six samples were allocated to raters three different ways: (1) single rater reading of all responses (SM1); (2) assignment of each c.r. response to a different rater (SM2); and (3) splitting the c.r. items into thirds, with a different rater for each portion (SM3). SSM1 readings produced average total c.r. scores that were greater than the average total of c.r. scores produced in the SM2 condition. Average total c.r. scores when students' responses were allocated to three different raters (SM3) were similar to those of the SM2 condition. Results suggest that for tests with relatively large numbers of c.r. items, the use of as few as three raters to score a student's examination could produce scores that were similar in magnitude and scale to those obtained by assigning a different rater to each item. (Contains 7 tables and 11 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 431 032

The Assignment of Raters to Items: Controlling for Rater Effects

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Robert Sykes

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Robert C. Sykes
CTB/McGraw-Hill

Mark Heidorn
Florida Department of Education

Guemin Lee
CTB/McGraw-Hill

TM029866

This paper was presented at the Annual Meeting of the National Council on Measurement in Education in Montreal, April 1999.

BEST COPY AVAILABLE

INTRODUCTION

The presence of more than one constructed response (c.r) item in an examination requires an allocation of readers (raters) to the various items. For paper-and-pencil examinations, especially those administered to large numbers of examinees, it is beneficial to minimize the movement of student responses (papers) within a pool of readers. The more raters involved in reading a student's set of c.r. responses the more movement of papers in the scoring center and hence the greater the chance of misplacing ratings and subsequently failing to incorporate a complete set of readings into a student's record.

With tests that call for multiple *examination* readings (i.e. having more than one reading of the complete set of c.r. item responses), the logistical task of effectively transferring papers is compounded by the number of additional examination readings. Each additional rater that reads a student's c.r. responses can be expected to slow and subsequently increase the cost of the scoring process because of the time required to allocate the appropriate papers. Hence, the most cost effective and efficient procedure for assigning readers to a student's examination is to assign one rater to read all of a student's c.r. responses (i.e. one rater per examination reading).

The use of a single rater for each examination reading (hereafter "single-rater-examination-reading" or single-rater-(e)reading) will expose all of a student's c.r. responses to a single rater's scoring accuracy for each item, however.

Significant differences in accuracy, as represented by differences in the degree of matching of an operational item rating with that obtained from an "expert" panel of judges (considered a true score: Sulsky & Balzer, 1988), have been found by Engelhard (1996) and others (McIntyre, Smith, & Hassett (1984).

In addition to item-specific characteristics, accuracy can be influenced by three different characteristics or response tendencies of the readers that span items (Saal, Downey, & Lahey (1980). "Central tendency" reflects the rater's reluctance to use either end of the scoring continuum. A strictness/leniency bias represents the tendency of a rater to provide ratings that are lower/higher than student performance warrants (Engelhard, 1994).

The third kind of across-item rater effect, halos, in which a rater is positively swayed by a student's response or responses to give more favorable ratings to other responses of the student, may be considered a type of leniency bias (Landy & Farr, 1980). It is difficult to efficiently conceal the fact that a rater has read other responses from the same student, and hence to forestall the potential for halo or "anti-halo" effects (the tendency for a previous response to reduce the score obtained on the following item) in the single-rater-(e)reading of a student's responses from a large-scale paper-and-pencil examination. Effects such as central tendency, severity/leniency biases, and halos/anti-halos can result in a restriction of the range of the

ratings.

The presence of one or more across-item rater effects would result in the accumulation of a particular rater's errors over a student's set of c.r. items under single-rater-(e)reading. This may be demonstrated through a simple modeling of a scored item response x_{ijk} as the sum of a student's true score t_{ij} for item i administered to student(person) j, a rater effect component at the item level δ_{ijk} that is unique to rater k but perhaps constant for subsets of the c.r. items, and an unique error component ε_{ijk} (that may contain other significant sources of variation such as items). The student's sum of c.r item scores or total c.r. score y_{jk} is then the sum of n item scores:

$$y_{jk} = \sum_{i=1}^{n} x_{ijk} = \sum_{i=1}^{n} (t_{ij} + \delta_{ijk} + \varepsilon_{ijk}) . \qquad (1)$$

The expected value of the total c.r. score upon repeated scorings by rater k is:

$$\mathrm{E}(y_{jk}) = \sum_{i=1}^{n} \mathrm{E}x_{ijk} = \sum_{i=1}^{n} (\mathrm{E}t_{ij} + \mathrm{E}\delta_{ijk}) = t_{j} + \sum_{i=1}^{n} \mathrm{E}\delta_{ijk} + \sum_{i=1}^{n} \mathrm{E}\varepsilon_{ijk} , \qquad (2)$$

where t_{j} is student j's true total c.r. score. The variance of student j's total c.r. score over repeated readings by rater k is:

$$\mathrm{var}(y_{jk}) = \mathrm{var}(\sum_{i=1}^{n} x_{ijk}) = \mathrm{var}\{t_{j} + \sum_{i=1}^{n} (\delta_{ijk} + \varepsilon_{ijk})\} = \sum_{i=1}^{n} \mathrm{var}(\delta_{ijk}) + \sum_{i=1}^{n} \mathrm{var}(\varepsilon_{ijk}) + 2\sum_{i}^{n} \sum_{<l}^{m} \mathrm{cov}(\delta_{ijk}, \delta_{ljk}) \qquad (3)$$

assuming that the ε_{ijk} are neither correlated across items or with

the δ_{ijk}. The summation in the last term extends over all values of i and l, from 1 to n, for which $i<l$.

The total c.r. score will deviate from the true c.r. score to the degree that the sum of the expected rater item effects or errors does not equal zero. This will occur if the sum of rater effects, such as halo or leniency, over a subset of items exceeds (or is less than) any other sum of effects, such as anti-halo effects or strictness, in the opposite direction. In the presence of substantial halo effects total c.r. scores would be inflated relative to true total c.r. scores.

The accumulation of rater effects would also impact the mean and variance of total c.r. scores. The mean total c.r. score of students from a sample of students under a single-rater-(e)reading would be:

$$\bar{y} = 1/N \sum_{j=1}^{N}\sum_{i=1}^{n} x_{ijk} = 1/N \sum_{j=1}^{N}\sum_{i=1}^{n}(t_{ij} + \delta_{ijk} + \varepsilon_{ijk}) = 1/N \sum_{j=1}^{N}(t_j + \sum_{i=1}^{n}\delta_{ijk} + \sum_{i=1}^{n}\varepsilon_{ijk})$$

$$= 1/N \sum_{j=1}^{N}(t_j + \sum_{i=1}^{n}\delta_{ijk}), \qquad (4)$$

assuming the sum of ε_{ijk} over items and students approximates 0 (i.e. has mean 0 in the population). The variance of the total c.r. score is then:

$$\mathrm{var}(y) = \mathrm{var}\sum_{i=1}^{n}(t_{ij} + \delta_{ijk} + \varepsilon_{ijk}) = \mathrm{var}(t_j) + \sum_{i=1}^{n}\mathrm{var}\,\delta_{ij} + \sum_{i=1}^{n}\mathrm{var}(\varepsilon_{ijk}) + 2\sum_{i}^{n}\sum_{<l}^{m}\mathrm{cov}(\delta_i, \delta_l), \qquad (5)$$

when the t_j's are independent of the δ_i's.

Consequently the mean total c.r. score of single-rater-(e)read students would differ from the mean of a sample of students having each c.r. item response read by a different reader (n-rater-(e)read) when $\sum_{i=1}^{n}\delta_i$ is not equal across the two types of reading. Under a n-rater-(e)reading procedure $\sum_{i=1}^{n}\delta_i$ would approximate 0 through the canceling of different raters' biases.

The variance of total c.r. scores obtained from single-rater-(e)reading may also be larger than the variance of n-rater-(e)read scores. This would occur because $2\sum_{i}^{n}\sum_{i<l}^{m}\mathrm{cov}(\delta_i,\delta_l)$ is likely to be larger for single-rater-(e)read scores. Rater error is more likely to be correlated within reader than across n different readers.

With the advent of imaging of c.r. item responses the logistical problem of allocating papers to readers is effectively solved. Each response of a student can be rated by a different reader. The expected value of the rater effects on the total c.r. score would then approach zero as the number of raters (c.r. items) "summed over" increases. However, while a different reader for each c.r. response would mimimize the error arising from a rater effect such as strictness/leniency bias, there are several reasons why it may be worthwhile to have fewer raters (but more than one) than the number of c.r. items.

First, the total c.r. score obtained for students evaluated by a reduced set of raters might contain less error variance than that of total scores of n-rater-(e)read students. Because fewer raters are summed over the sum of the variances of rater effects over items (the $\sum_{i=1}^{n} \mathrm{var}(\delta_{ijk})$ term in equation 3) may be less than that obtained through n-rater-(e)reads. Second, allowing a rater to read responses to more than one c.r. item might reduce the tedium of reading responses for only a single item, consequently helping to preserve reader attentiveness. (The latter advantage could also be accrued by maintaining separate readers for each of any individual student's responses but routing paper (images) such that a rater reads responses for more than one item from different students.)

The purpose of this research was to evaluate the effect of different modes or modalities of assigning raters to items. The impact on the total c.r. score, and subsequently total test score, of assigning a single versus multiple raters to an examination reading of a student's set of c.r. responses was evaluated for several mixed-item format tests.

METHOD

Instruments and Samples

Samples of approximately 2000 students were obtained for a Math field test form at each of Grades 5, 8, and 10 and for a Reading field test form at each of Grades 4, 8, and 10 of a large

state assessment. Each of the three Reading tests consisted primarily of items querying students about one of three or four literature passages.

In addition to large numbers of multiple choice (m.c.) items, the mixed-item format tests contained two types of c.r. items: two-point short response (s.r.) and four-point extended response (ex.r.) items. The number of scored items of each type are summarized below.

Content Area	Grade	# of Multiple Choice	# of Constructed Response S.R.(2 pt.)	Ex.R.(4 pt.)	Total # of C.R.pts.
Math	5	49	9	2	26
Math	8	52	9	2	26
Math	10	49	9	2	26
Reading	4	51	10	2	28
Reading	8	51	9	3	30
Reading	10	51	10	1	24

The forms were, on average, difficult for the field test population.

Rating Process

C.R. item responses for each of the selected students in the six samples were allocated to raters in three different ways or modalities. The first scoring modality (SM1) consisted of a single-rater-(e)reading of all of a student's c.r. responses. SM2 assigned each of a student's c.r. responses to a different rater (*n*-rater-(e)reading) while SM3 split the subset of c.r. items into approximate thirds, with a different rater assigned to each item block constituting approximately 1/3 of a student's

responses (three-rater-(e)reading). The incorporation of the third scoring modality allowed an evaluation of the potential for reducing or "averaging over" rater error with fewer raters per student than SM2. No rater participated in the scoring of more than one content area.

Modality-specific training and monitoring procedures (i.e. checksets and read-behinds) were implemented for each scoring modality. Subsamples of 30% of each of the six samples of student papers were submitted to a second examination reading under each modality. If the second item readings of the two-point s.r. items for students within these 30% "Multiple-Examination-Reading" subsamples of the complete samples did not agree exactly with the initial reading, a third reading was obtained. A third reading of a four-point ex.r. item was attained if the first two readings differed by more than one point.

Inter-rater reliability was evaluated for the participating pool of raters. Because reliability will appear greater when evaluated with samples containing students who did not respond to the c.r. items, reliability indices were obtained from the "Multiple-Examination-Reading" subsamples by trimming them of all students who obtained a 0 for a total c.r. score. Between 26 and 139 students were eliminated for this reason.

Agreement rates, both exact and approximate (within one point), and correlations across the first and second readings are presented for SM2 for the six grade/content trimmed subsamples in

Table 1. Exact agreement rates for the four-point ex.r. items in both Reading and Math are, as expected, generally lower than those obtained for the two-point s.r. items. Correlations between the first and second readings tend to be larger for the Math items.

Evaluation of Rater Effects

Total c.r. scores were computed by summing the c.r. item scores obtained within each of the three modalities for the single examination reading of all students in the six complete ("Single-Examination-Reading") samples. Means and standard deviations (sd's) of sets of the c.r. items, including total c.r. scores (hereafter total scores), were assessed across modalities. Means and sd's for each of the first two examination readings for the students in the "Multiple-Examination-Reading" subsamples were also evaluated.

Because each sampled student was scored in all three modalities, statistically powerful (in the sense of reduced error) within-subject comparisons could be evaluated for effects due to modality. Additionally, Generalizability and Decision studies were conducted to determine the reliability or consistency of both normative (G coefficient) and absolute (D coefficient) interpretations or classifications made on the basis of solely the c.r items.

RESULTS

"Single-Examination-Reading" Samples

Descriptive statistics for the total scores obtained within each of the scoring modalities for each of the six "Single-Examination-Reading" samples, ranging in size between 1,975 and 2,000 students, are provided in Table 2.

The mean total c.r. score for SM1 is notably greater than the means for SM2 and SM3 for the three Reading tests and for the Grade 5 Math form. SM1 means for Grade 8 and Grade 10 Math are very similar to the SM2 and SM3 means; the largest difference between the three pairs of means for Grade 10 Math is only .01. The sd's of SM1 total scores tend to be larger than the sd's for SM2 scores with the SM3 total sd's frequently falling between the sd's for the other two modalities.

The presence of students who did not attempt the c.r. items would attenuate differences due to scoring modality. Consequently the samples were trimmed of between 74 (Grade 8 Reading) and 552 (Grade 10 Math) students who obtained a total c.r. score of 0. Means and sd's for the trimmed "Single-Examination-Reading" samples are presented in Table 3a.

The difficulties (defined as the mean of SM1 scores divided by the total number of c.r. points) of the six sets of c.r. items (hereafter tests) using the trimmed samples range between .25 and .37. The Math tests are more difficult than the Reading. The total mean for SM1 exceeds the SM2 means with one exception, Grade 8 Math where both modality means equal 6.44, and are always

larger than the SM3 means. Sd's for the SM 1 scores are always larger than those for SM2 while the SM3 sd's frequently fall between those for the other two modalities.

Table 3b contains product moment correlations of the total c.r. scores across scoring modalities for the trimmed samples. The total scores tend to be highly correlated, with the smallest correlations occurring between SM1 versus SM2 and SM1 versus SM3 for Grade 10 Reading (.87 and .85, respectively). The correlations among the total scores for the three Math tests exceed the corresponding modality correlations for the other two Reading tests by .01 to as much as .06.

"Multiple-Examination-Reading' Subsamples

Representativeness of Trimmed Subsamples

Tables 4a and 4b contain scoring modality means and standard deviations, within and across the three item blocks, for the two examination readings (ERs) obtained for the trimmed "Multiple-Examination-Reading" Reading and Math subsamples, respectively. The *overall* (averaged over both examination readings) total means and sd's may be compared to the corresponding modality means for the trimmed "Single-Examination-Reading" samples in Table 3a to gauge the representativeness of the subsamples to their parent samples.

The trimmed overall Reading modality subsample means and sd's for Grade 4 and Grade 8 in Table 4a tend to be very similar to their corresponding sample statistics (e.g. an overall mean of 8.86 and sd of 4.96 for SM3 for the trimmed Grade 4 subsample in

Table 4a versus 8.92 and 5.05 for the trimmed sample). However, the trimmed subsample Grade 10 Reading means are roughly ½ point below the corresponding sample means.

The three overall modality means for each of the three Math subsamples in Table 4b are always less than or equal to one quarter of a score point below their corresponding trimmed "Single-Examination-Reading" modality means. The total (overall) scoring modality standard deviations for the trimmed Math subsamples are similar to their sample counterparts, varying unsubstantially above or below the corresponding sample sd's.

Comparisons Using Examination Readings

Comparability of Examination Readings

Total scores obtained through the second examination reading serve as a replication of those obtained from the first reading. A comparison of within-modality differences in ER means across the six tests indicate a range of insubstantial differences, varying between .00 for the two means for SM3 in Grade 8 Math (both 6.18 in Table 4b) to a .26 difference for the two SM2 means for Grade 5 Math (6.86 versus 7.12 for ER1 and ER2, respectively). Differences between ER total score *sd's* within modality tend to be small, with the largest difference being .11 for both SM1 for Grade 10 Reading (4.79 for ER1 versus 4.68 for ER2) and SM1 for Grade 10 Math (5.46 for ER1 versus 5.57).

Tests of Modality Differences

Comparisons of the 12 SM1 ER total score averages (two for each of the six grade/content area tests) with the 12 SM2 ER

averages indicates that, with the exception of Grade 8 Math, the SM1 means always exceed the SM2 means. The 12 SM3 ER means tend to be similar to the SM2 means.

Differences between SM1 versus SM2 total scores for the first and second examination readings (ER1:SM1-SM2 and ER2:SM1-SM2), as well as SM3 versus SM2 total scores for the two examination readings (ER1:SM3-SM2 and ER2:SM3-SM2) were evaluated for significance with t-tests. Because multiple significance tests were conducted a significance level of $p=.05/4=.0125$ was established for each of the four comparisons within a grade/content area. Asterisks denote in Tables 4a and 4b the significant comparisons.

All six SM1-SM2 mean differences for the three Reading tests were significant in favor of SM1 as compared to none of the six SM3-SM2 mean differences. The three Math tests varied in the significance of their SM1-SM2 differences: both mean ER total score differences were significantly positive for Grade 5, one ER mean score difference (ER2) was significantly positive for Grade 10, and neither were significant for Grade 8. One of the six Math SM3-SM2 comparisons was significant in favor of SM3, that for ER1 with Grade 5 (SM3:7.05, SM2:6.86). The other SM3-SM2 ER mean difference for Grade 5 Math was borderline, nonsignificantly *negative* in favor of SM2 ($p \leq .017$).

Although differences in sd's were not tested for significance, the two SM2 total ER sd's for each of the six grade/content areas were always smaller than the SM1 sd's for the

same test. With the exception of Grade 4 Reading, SM3 total score sd's were always smaller than the SM1 sd's and fell between the SM1 and SM2 sd's for all but that grade/content area and Grade 8 Math.

Sources of Modality Differences

Item Blocks

Tables 4a and 4b also portray ER means by item blocks, the partitions of approximately one third of a student's responses to the c.r. items read by a single rater under SM3. To the extent that the larger SM1 means (relative to those for SM2) for the three Reading tests and the Grades 5 and 10 Math tests are due to rater effects that accumulate successively over the c.r. items, SM1 item block (IB) means should progressively diverge from SM2 IB means. SM3 means would expectedly not demonstrate this divergence, relative to SM2 means, because of the use of a different rater to score a student's c.r. responses in each IB. A SM3 IB mean could substantially differ from a SM2 mean if rater effects had accumulated within the IB, however.

Patterns of increases in overall (averaged over ERs) SM1 IB means may be assessed against the pattern of non-increasing overall SM1 IB means seen for the Grade 8 Math test. The average overall SM1 versus SM2 means for IB 1 through IB 3 in Table 4b are: 3.90 versus 3.88, 1.52 versus 1.51, and .81 versus .86 for SM1 and SM2 in IB3. A marked contrast to the comparability demonstrated across the two scoring modalities for the Grade 8 Math IB means are the relative increases found for the Grade 8

Reading overall SM1 means. Average SM1 scores become increasingly larger than the SM2 means over IB's: 3.29 versus 3.06, 3.37 versus 3.11, and 5.00 versus 4.57, for differences of .23, .26, and .43, respectively. The other four grade/content areas demonstrate relative increases of SM1 means in some IB's with approximately equivalent SM1 and SM2 means in the other IB's.

The SM3 IB overall means are more similar to the SM2 overall means for the three Reading tests than are the SM1 means. They are not as distinctively similar to the SM2 means for the three Math tests because the Math SM1 IB means tend to demonstrate smaller (relative) increases.

Item Average Scores

In order to further delineate the nature of modality differences, average scores for the item constituents of the item blocks were computed. The average scores are presented in Tables 5a, 5b, and 5c for the three Reading tests and Tables 6a, 6b, and 6c for the three Math tests. Differences between item modality means, SM1-SM2 and SM3-SM2, that equal or exceed twice the standard error (s.e.) of the corresponding SM2 mean are bolded. Differences that were less than -2 times the s.e. of the SM2 mean are printed in bolded italics. The large number of comparisons caution against interpreting flagged means as significant at the nominal significance level ($p \approx .05$).

There are many fewer instances of significant positive or negative SM1 or SM3 mean differences from the baseline SM2 means

for Math than Reading when assessed against the criterion. There are 12 instances for Math (across examination readings) compared to 49 substantially deviant SM1 or SM3 means for the three Reading tests. Ten of the 12 substantial Math deviations are positive and there is a fairly even split between the number of significant SM1 and SM3 deviations (seven versus five). There are only two occurrences of adjacent significant deviations for the Math items (items #11 and #20 for ER1, Grade 5 and items #22 and #23 for ER2, Grade 8).

A very substantial portion of the differences between the Grade 5 Math *total* SM1 versus SM2 means for both ER1 and ER2 in Table 4b may be attributed to the significant positive SM1 deviation for item #11, a 4 point ex.r. item (SM1:1.84 - SM2:1.39 or .45 for ER1 and SM1:1.89 - SM2:1.58 for ER2).

Of the 49 significant SM1 and SM3 Reading deviations, substantially less than half (18) consist of positive or negative *SM3* deviations. Both SM1 and SM3 deviations for Reading occur more frequently adjacent to one another with some sets of SM3 adjacent deviations spanning item blocks, implying a continuation of substantial deviation over the substitution of a different rater reading the students' c.r. responses.

In addition to the runs of adjacent positive deviations (likely denoting halo effects), there are several instances of negative or attenuating effects. Perhaps the most interesting occurrences of negative effects are for the last two items in Grade 10 Reading (items #55 and #58 for both the ERs) and item #7

in the Grade 8 Reading test. The attenuating effect noted for items #55 and #58 may represent the effects of tedium or anti/halo effects. Item #7 falls between two substantial positive SM1 deviations (item #3 and #11) for ER1 of Grade 8 Reading. The item also has a substantial negative SM1 deviation for ER2 and similarly precedes a significant positive SM1 deviation for item #11.

Agreements in both the direction and significance of both SM1 and SM3 differences are common *across* ERs for the Reading tests (19 agreements in significant positive or negative deviations, 11 disagreements) but not the Math tests (three agreements, six disagreements). Furthermore, agreement in terms of direction is found in six of the 11 instances of disagreement for Reading. (In two of the instances of disagreement the nonsignificant SM mean equaled the corresponding SM2 mean.) An example of this is the SM1 mean for the Grade 8 Reading item #3 in ER2. As opposed to the item mean for ER1, the SM1 mean within ER2 is not significantly deviant by the criterion, although it does differ from the SM2 mean in the same positive direction (1.15 for SM1 versus 1.10 for SM2).

The consistent presence of groups of adjacent, significant SM1 mean differences across *Reading* ERs (with the possible exception of items #55 and #58 in Grade 10 Reading), as well as groups of significant SM3 differences *within* item blocks, is likely due to the passage-linked nature of the items. Consequently these deviations may be attributed to halo or anti-

halo effects, rather than the presence of more general strictness/leniency biases. The latter might be presumed to have a mean of 0 at the item level, with the number of strict raters approximately balanced by the number of lenient raters. Types of halo effects cannot, however, readily account either for significant SM1 or SM3 modality differences when they occur for the first scored c.r. item in the test or for significant SM3 differences when they occur for the first item in the second or third item block.

Generalizability and decision Studies

Modeling Components of Significant Variation

Generalizability (G) and decision (D) studies are commonly conducted to estimate the magnitude of individual sources of variation and predict the effect of adding levels of facets (effects) such as readers. The presence of a fixed effect due to scoring modality requires a generalization of the simple model used earlier to evaluate the potential for rater errors to accumulate over items. A more general model has an unreplicated item rating x_{mijk} as a combination of a fixed effect of scoring modality, τ_m, random effects attributable to an item π_i, student (person) β_j and rater δ_k, and interactions of the fixed and random effects:

$$x_{mijk} = \mu + \tau_m + \pi_i + \beta_j + \delta_k + \tau\pi_{mi} + \tau\beta_{mj} + \tau\delta_{mk} + \pi\beta_{ij} + \pi\delta_{ik} + \beta\delta_{jk}$$

$$+ \tau\pi\beta_{mij} + \tau\pi\delta m_{ik} + \tau\beta\delta_{mjk} + \pi\beta\delta_{ijk} + \varepsilon_{mijk}$$

If one or more of the effects are nested in a generalizability study not all variance components of the model can be independently estimated; some of them are confounded with others.

The presence of both fixed and random effects requires that a mixed model methodology be utilized for the simultaneous estimation of both types of effects. Unfortunately more than one version of the general model are required for item responses scored under the three scoring modalities.

The generalizability study designs for the three scoring modalities are as follows:

Modality	Design
1	(person:rater) x item
2	person x (rater:item)
3	person x (rater:item) partially nested

where "x" denotes a crossing of the levels of the adjacent effects or facets and ":" indicates the effect on the left is nested within levels of the effect on the right. The second and third modality designs share the nesting of raters within items while SM1 or the single-rater-(e)readings, have persons nested within raters. SM3, however, has raters nested within item blocks at the same time persons are nested within raters within item blocks. (Hence, it is not possible to simply characterize the design.)

Consequently some terms are estimable in one of the modality models but not in the other. For example, the item-by-rater

interaction $\pi\delta_{ik}$ is estimable for SM1 but confounded with other terms in the model for SM2 while, conversely, the item-by-person interaction $\pi\beta_{ij}$ is estimable for SM2 but confounded for the SM1 model. The three-way item-by-person-by-rater interaction $\pi\beta\delta_{ijk}$ is not estimable under the SM1 or SM2 models.

The existence of three different facet designs prevents the use of a mixed model methodology, such as the SAS PROC MIXED procedure (1997), to simultaneously estimate both fixed and random effects. If the rater effect and all interactions involving raters could be assumed insignificant the rater terms could be dropped from the two different modality models, resulting in a common, estimable mixed model.

Nonsubstantial rater effects or interactions may be questioned, however, given the differences in means and sd's for SM1 versus SM2 and SM3 previously described. The presence of halo effects in the SM1 scorings, as well as possibly to a smaller degree the SM3 scorings within item blocks, for the three Reading tests and the Grade 5 or Grade 10 Math tests could imply the presence of a nonzero SM-by-person-by-rater interaction.

On the other hand deviant SM1 or SM3 item averages for the Reading tests (in Tables 5a, 5b, and 5c) that are consistent across two examination readings, and hence two different sets of readers, may portend a substantial SM-by-item-by-person rather than SM-by-person-rater interaction. This would imply that rater errors could be commonly induced by item characteristics,

specifically their linkage to Reading passages.

Because interactions with raters could not be estimated in a common model that excluded this effect, any significant variation due to these interactions would be confounded with other terms in the residual. Consequently the power of a test of the main effect of scoring modality using the residual as the error term would be reduced.

Within-Modality Analyses

As a means to further define the particular sources of variation in the scored item responses, G and D studies were conducted within the SM1 and SM2 scoring modalities. A procedure for estimating the variance components for the partially nested design of SM3 can not be captured by a single G study design, and consequently it was not included in the within-modality generalizability analyses.

Comparisons of the similarity of estimated variance components, including the residual, across the two modalities could provide clues to the significance of unmodeled interactions, including those involving scoring modality. Predictions of the effect of adding readers on the reliability of relative and absolute decisions could also be made within SM1 and SM2 through the estimation of G coefficients and index of dependability (ϕ coefficients).

The work of Brennan (1995) was used to estimate the two reliability indices for a relatively rare SM1 design that includes the object of measurement, persons, nested within

raters. Estimation of the G and ϕ coefficients for the SM2 design having persons crossed with raters nested within items was conducted in a manner specified by Shavelson and Webb (1991).

The generalizability of inferences both within and across modalities, made on the basis of within-modality estimated variance components and reliability indices, depends upon the extent that unmodeled effects, most notably modality, impact the relative or absolute standings of the scores. Previously described results indicate that the single-rater-(e)scoring of SM1 does influence both the dispersion and level of item and total scores.

Tables 7a and 7b contain MIVQUE0 estimates of variance components from the SAS VARCOMP procedure (SAS, 1988) for SM1 and SM2 for all first examination readings of the trimmed "Multiple-Examination-Reading" subsamples for Reading and Math, respectively. Turning first to the SM2 variance components, the item-by-person interaction is the largest source of variation across all six grade/content areas, constituting between 41.8% (Grade 4 Math) and 54.3% (Grade 5 Math) of total variation. The random item and person effects are the next largest sources of variation, with the magnitude of the residual term rivaling that of the former effects in Reading only.

G coefficients for a single examination reading under SM2, utilizing approximately two raters(n_r) to read all the student responses for each item, ranged between .768 (Grade 10 Reading) and .837 (Grade 8 Reading). A doubling of the number of readers,

producing the effect of averaging over two examination readings, results in very small gains in both relative and absolute SM2 reliabilites (increases less than .02). These relatively small increases in reliability are comparable to the modest effects of adding raters noted by Linn & Burton (1994).

Single examination readings under SM1 result in lower relative and absolute reliabilities, when compared against the corresponding single-reading reliabilities for SM2, for all tests but Grade 10 Reading and Grade 8 Math. Both types of reliabilities for the latter test are very similar across modalities, differing by at most .009 (A ϕ coefficient of .714 for SM1 versus .705 for SM2). It is difficult to interpret the substantiveness of the larger reliability coefficients for SM1 for Grade 10 Reading because of the unaccounted effects associated with SM1 scoring. Grade 10 Reading demonstrates the largest difference in SM1 versus SM2 total scores (Table 4a: 7.22(SM1) - 6.54(SM2) = .68) and the largest number of significant item deviations for SM1 (Table 5c).

The addition of a second examination reading under SM1 does not increase the reliability of total scores for any of the six tests, unlike the very modest increases noted for SM2. This is because neither the $\hat{\sigma}^2_{\pi\delta}$ or $\hat{\sigma}^2_{\varepsilon}$ terms constituting relative variance nor the $\hat{\sigma}^2_{\pi}$ term added to complete the absolute variance is reduced by adding raters.

DISCUSSION/CONCLUSIONS

Increases in rater assigned item scores due to differences in the mode or modality of scoring the c.r. items of mixed-item format tests resulted in substantive increases in group averages on the total c.r. component score for five of the six Reading and Math tests assessed. Single-rater-(e)reading of a student's complete set of c.r. responses produced average total c.r. scores that were .23 to .68 (between approximately 5% and 15% of a total c.r. score sd) greater than the average total c.r. score obtained for the same large samples of students when a different rater scored each of the 11 to 12 two point and four point c.r. items (*n*-rater-(e)reading). Average total c.r. scores for these students when each student's c.r. responses were allocated to three different raters, or three-rater-(e)read, were very similar to the averages obtained with *n*-rater-(e)reading.

The dispersion of total c.r. scores were increased under single-rater-(e)reading compared to both n-rater or three-rater-(e)reading. Dispersions for three-rater-(e)reading total c.r. scores were also increased, relative to n-rater-(e)readings, although to a lesser degree than for single-rater-(e)readings. Both the increase in level and dispersion of total c.r. scores attained through single rater scoring are predicted by models that allow for an accumulation of rater errors over the set of scored c.r. item responses.

The generally larger increases found for the single-rater-(e)readings for the three Reading tests could be linked to

increases in ratings for a number of individual items, frequently occurring within sets of adjacent items. Increased ratings for three-rater-(e)read items occurred to a lesser extent and less frequently within sets of adjacent items. Both the greater incidence of sets of adjacent items with increased average scores and their frequent, consistent presence in two separate examination readings supports attributing the increases to the passage-linked nature of the Reading items. Work is needed to describe the particular relationship among items within the passages and the manner in which they may influence ratings.

Increased average item scores for sets of adjacent items supports a causative role for halo effects in the inflation of scores. The great difficulty of concealing from a reader the source of previously read responses makes it likely that halo effects are present in the scores obtained from single-rater-(e)scoring of large-scale paper-and-pencil tests.

The occurrence of significantly increased ratings, under either single-rater or three-rater-(e)reading, for several "first-scored" items cannot be attributed to halo effects arising from exposure to the student's previous response, however. Some average item scores that were significantly less than those obtained under n-rater-(e)reading suggests the presence of anti-halo as well as halo effects on rater judgments through the course of scoring an examination. If additional work determines the same judges can demonstrate both effects over items it would suggest that rater behavior may not be sufficiently modelled by

fitting a single rater strictness/leniency parameter.

It was not possible to fit a mixed model to scored item responses from all three scoring modalities because it could not be established at this time that all interactions involving raters were insubstantial. The inability to fit a general mixed model prevented estimating the degree that reliability was attenuated by the use of single rater scoring.

Generalizability and D studies could be conducted within two of the three modalities (SM1 and SM2) that had facet designs for which variance components could be estimated. The very modest improvement in the reliability of relative or absolute classification decisions that is obtained by adding an additional examination reading under SM2 is consistent with previous research.

Additional work in more specifically characterizing sources of variation in item scores may allow a more direct comparison of the reliability of single-rater versus multiple-rater-(e)readings through the fitting of a general, mixed model.. The results of the present study suggest that for tests with relatively large numbers of c.r. items the use of as few as three raters to score a student's examination could produce scores that were similar (in magnitude and scale) to those obtained by assigning a different rater to each item.

References

Brennan, R. L. (1995). The conventional wisdom about group mean Scores. *Journal of Educational Measurement*, 32, 385-396.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93-112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*, 56-70.

Landy, R.J., & Farr, J.L. (1980). Performance rating. *Psychological Bulletin*, *87*, 72-107.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational measurement: Issues and Practice*, *13*, 5-8.

McIntyre, R. M., Smith, D. E.,& Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, *69*, 147-156.

Saal, F. E., Downey, R. G.,& Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413-428.

SAS Institute. (1988). *SAS/STAT user's guide*. Cary, NC:Author.

SAS Institute. (1997). *SAS/STAT software:Changes and enhancements through release 6.12*. Cary, NC:Author.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

Sulsky, L. M.,& Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*, 497-506.

Table 1
Inter-rater Reliability Statistics
"Multiple-Examination-Reading" Subsamples: Scoring Modality 2
(Excludes Total CR Scores of 0)

Reading

Grade 4
(n=630)

Item #	Pt Value	Agreement Rate			Corr
		Exact	Approximate (within 1 pt.)	Total (Exact + Approx.)	
8	2	0.90	0.10	1.00	0.90
9	2	0.77	0.22	0.99	0.78
14	2	0.79	0.20	0.99	0.77
17	4	0.81	0.18	0.99	0.89
25	2	0.87	0.13	1.00	0.80
30	2	0.85	0.15	1.00	0.71
34	2	0.82	0.18	1.00	0.77
39	2	0.84	0.16	1.00	0.72
49	2	0.90	0.09	1.00	0.91
52	2	0.85	0.15	1.00	0.74
54	4	0.61	0.37	0.97	0.74

Grade 8
(n=600)

Item #	Pt Value	Agreement Rate			Corr
		Exact	Approximate (within 1 pt.)	Total (Exact + Approx.)	
3	2	0.73	0.26	0.98	0.71
7	2	0.80	0.20	0.99	0.79
11	4	0.69	0.29	0.98	0.81
16	2	0.83	0.16	0.99	0.78
19	2	0.71	0.28	0.99	0.61
29	4	0.73	0.25	0.98	0.83
34	2	0.83	0.15	0.98	0.69
37	2	0.84	0.15	0.99	0.66
47	2	0.95	0.05	1.00	0.91
50	2	0.95	0.05	1.00	0.96
54	4	0.67	0.28	0.95	0.88
57	2	0.94	0.07	1.00	0.93

Grade 10
(n=553)

Item #	Pt Value	Agreement Rate			Corr
		Exact	Approximate (within 1 pt.)	Total (Exact + Approx.)	
8	2	0.81	0.19	1.00	0.78
15	2	0.73	0.25	0.98	0.69
19	2	0.76	0.24	0.99	0.52
21	2	0.70	0.29	0.99	0.57
31	2	0.89	0.11	1.00	0.73
37	2	0.85	0.13	0.99	0.87
42	2	0.95	0.05	1.00	0.95
50	2	0.97	0.03	1.00	0.92
53	4	0.66	0.31	0.97	0.87
55	2	0.86	0.13	0.99	0.79
58	2	0.85	0.15	1.00	0.75

Mathematics

Grade 5
(n=561)

Item #	Pt Value	Agreement Rate			Corr
		Exact	Approximate (within 1 pt.)	Total (Exact + Approx.)	
10	2	0.97	0.03	1.00	0.97
11	4	0.53	0.34	0.87	0.68
20	2	0.91	0.09	1.00	0.82
21	2	0.93	0.06	1.00	0.88
22	2	0.94	0.05	0.99	0.95
41	2	0.94	0.06	1.00	0.93
42	2	0.96	0.03	0.99	0.96
43	2	0.92	0.08	1.00	0.93
51	4	0.77	0.19	0.96	0.89
52	2	0.93	0.07	1.00	0.94
53	2	0.99	0.01	1.00	0.99

Grade 8
(n=564)

Item #	Pt Value	Agreement Rate			Corr
		Exact	Approximate (within 1 pt.)	Total (Exact + Approx.)	
11	4	0.81	0.18	0.98	0.90
12	2	0.96	0.04	1.00	0.93
13	2	0.95	0.05	1.00	0.96
22	2	0.95	0.04	0.99	0.92
23	2	0.89	0.11	1.00	0.83
24	2	0.92	0.07	1.00	0.92
42	2	0.98	0.02	1.00	0.98
43	2	0.78	0.22	0.99	0.64
54	2	0.94	0.06	1.00	0.94
55	2	0.99	0.01	1.00	0.98
56	4	0.90	0.10	0.99	0.89

Grade 10
(n=507)

Item #	Pt Value	Agreement Rate			Corr
		Exact	Approximate (within 1 pt.)	Total (Exact + Approx.)	
9	2	0.89	0.10	0.99	0.92
10	2	0.84	0.13	0.97	0.74
11	4	0.86	0.13	0.99	0.96
19	2	0.98	0.01	1.00	0.98
20	2	0.93	0.07	1.00	0.84
21	2	0.99	0.01	1.00	0.99
40	2	0.90	0.07	0.98	0.84
41	2	0.97	0.03	1.00	0.97
48	2	0.93	0.07	1.00	0.92
49	2	0.98	0.02	1.00	0.98
50	4	0.95	0.04	0.99	0.97

BEST COPY AVAILABLE

Table 2
Descriptive Statistics for Total CR Scores
"Single-Examination-Reading" Samples

Grade	Content Area	Form	Total # of C.R. pts.	# of ER Items	N	Scoring Modality					
						1		2		3	
						Mean	SD	Mean	SD	Mean	SD
4	Reading	A	28	2	1999	8.79	5.32	8.43	5.02	8.36	5.34
8	Reading	D	30	3	1975	10.62	6.06	10.19	5.81	10.23	5.91
10	Reading	D	24	1	2000	6.56	5.25	5.85	4.73	5.89	4.85
5	Math	C	26	2	1996	6.50	5.35	6.00	5.17	6.12	5.21
8	Math	C	26	2	1987	5.30	4.89	5.29	4.89	5.25	4.88
10	Math	B	26	2	2000	5.11	5.54	5.11	5.44	5.12	5.48

Table 3a
Descriptive Statistics for Total CR Scores
"Single-Examination-Reading" Samples
(Excludes Total CR Scores of 0)

Grade	Content Area	Form	Total # of CR pts.	# of ER Items	Difficulty [Mean (SM1)/ # CR pts.]	N	Scoring Modality					
							1		2		3	
							Mean	SD	Mean	SD	Mean	SD
4	Reading	A	28	2	0.33	1873	9.36	5.00	8.98	4.00	8.92	5.05
8	Reading	D	30	3	0.37	1901	11.03	5.81	10.57	5.59	10.62	5.69
10	Reading	D	24	1	0.33	1652	7.86	4.85	7.02	4.37	7.06	4.52
5	Math	C	26	2	0.30	1668	7.69	5.07	7.14	4.90	7.27	4.93
8	Math	C	26	2	0.25	1624	6.44	4.70	6.44	4.69	6.38	4.69
10	Math	B	26	2	0.27	1448	6.97	5.45	6.80	5.30	6.80	5.35

Table 3b
Product Moment Correlations
"Single-Examination-Reading" Samples
(Excludes Total CR Scores of 0)

Reading

Grade 4

Scoring Modality	Scoring Modality		
	1	2	3
1	1.00	0.95	0.94
2		1.00	0.94
3			1.00

Grade 8

Scoring Modality	Scoring Modality		
	1	2	3
1	1.00	0.94	0.92
2		1.00	0.94
3			1.00

Grade 10

Scoring Modality	Scoring Modality		
	1	2	3
1	1.00	0.87	0.85
2		1.00	0.92
3			1.00

Mathematics

Grade 5

Scoring Modality	Scoring Modality		
	1	2	3
1	1.00	0.96	0.95
2		1.00	0.96
3			1.00

Grade 8

Scoring Modality	Scoring Modality		
	1	2	3
1	1.00	0.97	0.96
2		1.00	0.97
3			1.00

Grade 10

Scoring Modality	Scoring Modality		
	1	2	3
1	1.00	0.97	0.97
2		1.00	0.98
3			1.00

BEST COPY AVAILABLE

Table 4a
Average Reading Scores by Scoring Modality, Item Block, and Examination Reading
"Multiple-Examination-Reading" Subsamples
(Excludes Total CR Scores of 0)

Grade	Content Area	Form	Total # CR pts.	# of ER Items	N	Scoring Modality	Examination Reading (ER)	Item Block 1 Mean		Item Block 2 Mean		Item Block 3 Mean		Total (1+2+3) Mean		Total (1+2+3) SD	
								ER	Overall	ER	Overall	ER	Overall	ER	Overall	ER	Overall
4	Reading	A	28	2	630	1	ER 1	3.79	3.79	2.90	2.90	2.67	2.66	9.36 *	9.35	5.04	4.97
							ER 2	3.79		2.91		2.65		9.35 *		5.03	
						2	ER 1	3.63	3.60	2.86	2.84	2.41	2.40	8.90	8.84	4.71	4.68
							ER 2	3.56		2.83		2.39		8.88		4.79	
						3	ER 1	3.61	3.62	2.71	2.73	2.53	2.51	8.84	8.86	5.09	4.96
							ER 2	3.61		2.74		2.50		8.88		5.01	

Grade	Content Area	Form	Total # CR pts.	# of ER Items	N	Scoring Modality	Examination Reading (ER)	Item Block 1 Mean		Item Block 2 Mean		Item Block 3 Mean		Total (1+2+3) Mean		Total (1+2+3) SD	
								ER	Overall	ER	Overall	ER	Overall	ER	Overall	ER	Overall
8	Reading	D	30	3	600	1	ER 1	3.24	3.29	3.39	3.37	4.43	5.00	11.06 *	11.10	5.91	5.81
							ER 2	3.34		3.35		4.46		11.15 *		5.92	
						2	ER 1	3.01	3.06	3.11	3.11	4.58	4.57	10.70	10.74	5.70	5.67
							ER 2	3.12		3.12		4.55		10.78		5.78	
						3	ER 1	3.15	3.15	3.16	3.16	4.42	4.42	10.74	10.73	5.83	5.71
							ER 2	3.14		3.15		4.43		10.73		5.81	

Grade	Content Area	Form	Total # CR pts.	# of ER Items	N	Scoring Modality	Examination Reading (ER)	Item Block 1 Mean		Item Block 2 Mean		Item Block 3 Mean		Total (1+2+3) Mean		Total (1+2+3) SD	
								ER	Overall	ER	Overall	ER	Overall	ER	Overall	ER	Overall
10	Reading	D	24	1	553	1	ER 1	2.43	2.41	2.60	2.57	2.26	2.24	7.30 *	7.22	4.79	4.53
							ER 2	2.39		2.54		2.21		7.15 *		4.68	
						2	ER 1	2.09	2.07	2.14	2.14	2.29	2.33	6.52	6.54	4.22	4.17
							ER 2	2.05		2.14		2.36		6.55		4.24	
						3	ER 1	2.18	2.22	2.25	2.23	2.09	2.14	6.52	6.60	4.36	4.30
							ER 2	2.26		2.21		2.20		6.67		4.46	

* Indicates significant difference in mean score relative to corresponding examination reading for Scoring Modality 2 $p \leq .0125$

BEST COPY AVAILABLE

Table 4b
Average Mathematics Scores by Scoring Modality, Item Block, and Examination Reading
"Multiple-Examination-Reading" Subsamples
(Excludes Total CR Scores of 0)

Grade	Content Area	Form	Total # CR pts	# of ER Items	N	Scoring Modality	Examination Reading (ER)	Item Block 1 Mean		Item Block 1 Mean		Item Block 1 Mean		Total (1+2+3) Mean		Total (1+2+3) SD	
								ER	Overall	ER	Overall	ER	Overall	ER	Overall	ER	Overall
5	Math	C	26	2	561	1	ER 1	3.40	3.43	2.39	2.38	1.67	1.67	7.46 *	7.48	5.17	5.13
							ER 2	3.45		2.38		1.67		7.50 *		5.20	
						2	ER 1	2.89	2.97	2.35	2.37	1.63	1.66	6.86	6.99	4.99	4.97
							ER 2	3.06		2.38		1.68		7.12		5.04	
						3	ER 1	3.05	3.04	2.41	2.40	1.59	1.58	7.05 *	7.02	5.04	4.99
							ER 2	3.02		2.40		1.56		6.98		5.05	

Grade	Content Area	Form	Total # CR pts	# of ER Items	N	Scoring Modality	Examination Reading (ER)	Item Block 1 Mean		Item Block 1 Mean		Item Block 1 Mean		Total (1+2+3) Mean		Total (1+2+3) SD	
								ER	Overall	ER	Overall	ER	Overall	ER	Overall	ER	Overall
8	Math	C	26	2	564	1	ER 1	3.86	3.90	1.52	1.52	0.82	0.81	6.20	6.23	4.55	4.51
							ER 2	3.95		1.51		0.80		6.25		4.56	
						2	ER 1	3.88	3.88	1.53	1.51	0.86	0.86	6.27	6.24	4.49	4.42
							ER 2	3.87		1.50		0.85		6.22		4.41	
						3	ER 1	3.86	3.85	1.50	1.52	0.82	0.81	6.18	6.18	4.48	4.42
							ER 2	3.84		1.53		0.81		6.18		4.47	

Grade	Content Area	Form	Total # CR pts	# of ER Items	N	Scoring Modality	Examination Reading (ER)	Item Block 1 Mean		Item Block 1 Mean		Item Block 1 Mean		Total (1+2+3) Mean		Total (1+2+3) SD	
								ER	Overall	ER	Overall	ER	Overall	ER	Overall	ER	Overall
10	Math	B	26	2	507	1	ER 1	3.46	3.55	1.45	1.46	1.79	1.79	6.70	6.78	5.46	5.45
							ER 2	3.61		1.47		1.79		6.87 *		5.57	
						2	ER 1	3.44	3.41	1.32	1.31	1.81	1.81	6.58	6.55	5.37	5.32
							ER 2	3.39		1.33		1.81		6.51		5.32	
						3	ER 1	3.48	3.48	1.41	1.39	1.80	1.78	6.69	6.65	5.42	5.39
							ER 2	3.49		1.37		1.76		6.62		5.42	

* Indicates significant difference in mean score relative to corresponding examination reading for Scoring Modality 2. $p \leq .0125$.

BEST COPY AVAILABLE

Table 5a
Grade 4 Reading Average Item Scores by Item Block
(Excludes Total CR Scores of 0)

(n=561)

Examination Reading 1

Item #	Pt. Value	Scoring Modality 1 Mean	1 SD	2 Mean	2 SD	3 Mean	3 SD
6	2	0.47	0.79	0.44	0.76	0.47	0.78
9	2	0.92	0.75	0.90	0.77	0.91	0.75
14	2	0.65	0.76	0.64	0.69	0.65	0.78
17	4	**1.74**	1.03	1.65	1.04	1.59	1.04
25	2	0.57	0.62	0.56	0.58	0.52	0.62
30	2	0.71	0.56	0.73	0.51	0.71	0.57
34	2	0.84	0.69	0.82	0.63	0.79	0.64
39	2	0.78	0.57	0.75	0.52	***0.69***	0.61
49	2	0.59	0.80	0.59	0.79	0.58	0.80
52	2	0.51	0.55	0.55	0.54	0.56	0.58
54	4	**1.06**	0.99	0.95	0.95	0.99	1.02
63	2	**0.51**	0.69	0.32	0.56	**0.39**	0.62

Examination Reading 2

Item #	Pt. Value	Scoring Modality 1 Mean	1 SD	2 Mean	2 SD	3 Mean	3 SD
6	2	0.47	0.79	0.44	0.76	0.49	0.79
9	2	0.92	0.76	0.88	0.76	0.92	0.75
14	2	0.66	0.76	0.62	0.74	0.63	0.78
17	4	**1.74**	1.06	1.62	1.04	1.59	1.01
25	2	0.55	0.60	0.54	0.58	0.53	0.62
30	2	0.72	0.55	0.74	0.52	0.71	0.56
34	2	0.86	0.71	0.81	0.63	0.79	0.63
39	2	0.78	0.55	0.74	0.53	0.71	0.61
49	2	0.59	0.81	0.58	0.78	0.56	0.79
52	2	0.50	0.55	0.54	0.54	0.55	0.57
54	4	**1.04**	0.98	0.95	0.96	1.00	0.99
63	2	**0.52**	0.67	0.32	0.58	**0.39**	0.62

Table 5b
Grade 8 Reading Average Item Scores by Item Block
(Excludes Total CR Scores of 0)

(n=600)

Examination Reading 1

Item #	Pt. Value	Scoring Modality 1 Mean	1 SD	2 Mean	2 SD	3 Mean	3 SD
3	2	**1.13**	0.79	1.05	0.77	1.10	0.74
7	2	***0.48***	0.63	0.57	0.72	0.60	0.72
11	4	**1.34**	1.14	1.12	0.97	1.13	1.00
37*	2	0.29	0.55	0.26	0.52	0.29	0.57
16	2	**0.82**	0.67	0.70	0.65	0.65	0.65
19	2	**0.97**	0.70	0.79	0.63	0.79	0.66
29	4	1.31	1.16	1.32	0.97	1.37	0.98
34	2	0.29	0.52	0.31	0.62	0.30	0.62
47	2	1.17	0.57	1.21	0.56	1.21	0.56
50	2	0.94	0.77	0.94	0.78	0.93	0.77
54	4	1.60	1.45	1.67	1.51	1.66	1.51
57	2	0.72	0.63	0.76	0.68	0.75	0.68

Examination Reading 2

Item #	Pt. Value	Scoring Modality 1 Mean	1 SD	2 Mean	2 SD	3 Mean	3 SD
3	2	1.15	0.79	1.10	0.74	1.09	0.78
7	2	***0.48***	0.62	0.60	0.72	0.61	0.71
11	4	**1.41**	1.13	1.13	1.00	***1.05***	0.97
37*	2	0.30	0.55	0.29	0.57	**0.39**	0.61
16	2	**0.80**	0.67	0.65	0.65	**0.75**	0.66
19	2	**0.95**	0.68	0.79	0.66	**0.87**	0.71
29	4	1.31	1.08	1.37	0.98	***1.22***	1.05
34	2	0.29	0.54	0.30	0.62	0.32	0.66
47	2	1.19	0.56	1.21	0.56	1.22	0.57
50	2	0.93	0.77	0.93	0.77	0.92	0.77
54	4	1.65	1.47	1.66	1.51	1.57	1.52
57	2	***0.70***	0.63	0.75	0.68	0.73	0.66

Table 5c
Grade 10 Reading Average Item Scores by Item Block
(Excludes Total CR Scores of 0)

(n=553)

Examination Reading 1

Item #	Pt. Value	Scoring Modality 1 Mean	1 SD	2 Mean	2 SD	3 Mean	3 SD
8	2	**0.69**	0.74	0.51	0.69	0.48	0.66
15	2	0.46	0.73	0.49	0.73	***0.37***	0.70
19	2	**0.41**	0.67	0.31	0.51	0.31	0.58
21	2	**0.87**	0.76	0.78	0.61	**1.03**	0.75
31	2	**0.58**	0.72	0.20	0.48	**0.30**	0.59
37	2	**0.90**	0.84	0.81	0.86	0.83	0.88
42	2	1.13	0.78	1.13	0.75	1.12	0.76
50	2	**0.35**	0.63	0.28	0.52	0.29	0.54
53	4	1.41	1.26	1.35	1.26	1.26	1.19
55	2	***0.27***	0.59	0.37	0.62	***0.27***	0.55
58	2	***0.23***	0.51	0.30	0.54	0.26	0.54

Examination Reading 2

Item #	Pt. Value	Scoring Modality 1 Mean	1 SD	2 Mean	2 SD	3 Mean	3 SD
8	2	**0.70**	0.74	0.50	0.66	0.49	0.66
15	2	0.43	0.71	0.46	0.74	***0.40***	0.72
19	2	**0.39**	0.67	0.29	0.53	**0.34**	0.61
21	2	**0.87**	0.76	0.80	0.60	**1.03**	0.75
31	2	**0.56**	0.73	0.21	0.48	**0.30**	0.62
37	2	0.85	0.81	0.82	0.89	0.81	0.87
42	2	1.13	0.79	1.11	0.76	1.09	0.76
50	2	**0.33**	0.61	0.28	0.52	0.30	0.55
53	4	1.38	1.25	1.41	1.30	***1.33***	1.22
55	2	***0.28***	0.59	0.36	0.62	***0.30***	0.56
58	2	***0.22***	0.51	0.31	0.56	0.27	0.52

Bolded values indicate differences in means SM1-SM2 or SM3-SM2 (2(SE) of SM2 mean

Bolded italicized values indicate differences in means SM1-SM2 or SM3-SM2 <| -2(SE)] of SM2 mean

*Discrete Item 37 in Grade 8 scored after Item 11 in order to preserve a 4-item-block

BEST COPY AVAILABLE

Table 6a
Grade 5 Mathematics Average Item Scores by Item Block
(Excludes Total CR Scores of 0)
(n=630)

Examination Reading 1

Item #	Pt. Value	Scoring Modality					
		1		2		3	
		Mean	SD	Mean	SD	Mean	SD
10	2	0.53	0.74	0.52	0.74	0.52	0.73
11	4	**1.84**	1.25	1.39	1.15	**1.56**	1.26
20	2	**0.29**	0.56	0.25	0.53	0.22	0.48
21	2	0.74	0.58	0.73	0.57	0.75	0.59
22	2	0.81	0.83	0.78	0.83	0.83	0.84
41	2	0.33	0.69	0.29	0.62	0.30	0.65
42	2	0.59	0.90	0.60	0.90	0.59	0.90
43	2	0.65	0.73	0.68	0.76	0.69	0.78
54	4	1.14	1.46	1.07	1.39	1.06	1.40
55	2	0.42	0.74	0.44	0.77	0.41	0.74
56	2	0.12	0.44	0.12	0.44	0.12	0.41

Examination Reading 2

Item #	Pt. Value	Scoring Modality					
		1		2		3	
		Mean	SD	Mean	SD	Mean	SD
10	2	0.53	0.74	0.51	0.73	0.53	0.74
11	4	**1.89**	1.27	1.58	1.25	1.49	1.24
20	2	0.29	0.56	0.25	0.53	0.25	0.51
21	2	0.75	0.58	0.71	0.56	0.75	0.58
22	2	0.81	0.83	0.81	0.84	0.81	0.84
41	2	0.34	0.69	0.31	0.67	0.29	0.65
42	2	0.58	0.89	0.59	0.89	0.59	0.89
43	2	0.65	0.73	0.67	0.75	0.71	0.78
54	4	1.14	1.47	1.12	1.41	1.07	1.43
55	2	0.41	0.73	0.44	0.75	0.39	0.70
56	2	0.12	0.42	0.12	0.43	0.10	0.37

Table 6b
Grade 8 Mathematics Average Item Scores by Item Block
(Excludes Total CR Scores of 0)
(n=564)

Examination Reading 1

Item #	Pt. Value	Scoring Modality					
		1		2		3	
		Mean	SD	Mean	SD	Mean	SD
11	4	1.87	1.16	1.90	1.14	1.92	1.16
12	2	0.71	0.57	0.72	0.59	0.72	0.58
13	2	0.84	0.78	0.82	0.76	0.82	0.79
22	2	0.44	0.76	0.44	0.77	0.39	0.71
23	2	0.31	0.54	0.34	0.59	**0.42**	0.61
24	2	0.40	0.74	0.44	0.76	0.38	0.71
42	2	0.26	0.64	0.27	0.64	0.27	0.64
43	2	**0.55**	0.56	0.48	0.56	0.43	0.60
54	2	0.48	0.70	0.49	0.70	0.48	0.69
55	2	0.12	0.44	0.11	0.44	0.11	0.43
56	4	0.22	0.69	0.26	0.72	0.23	0.75

Examination Reading 2

Item #	Pt. Value	Scoring Modality					
		1		2		3	
		Mean	SD	Mean	SD	Mean	SD
11	4	1.92	1.16	1.89	1.12	1.90	1.12
12	2	0.71	0.57	0.71	0.58	0.71	0.56
13	2	0.86	0.80	0.82	0.76	0.84	0.79
22	2	0.46	0.77	0.46	0.79	***0.39***	0.73
23	2	0.31	0.53	0.32	0.55	**0.41**	0.62
24	2	0.41	0.73	0.41	0.74	0.42	0.76
42	2	0.27	0.64	0.26	0.63	0.27	0.64
43	2	0.52	0.56	0.50	0.59	***0.44***	0.60
54	2	0.49	0.71	0.50	0.70	0.48	0.69
55	2	0.11	0.43	0.12	0.44	0.10	0.42
56	4	0.20	0.68	0.24	0.71	0.23	0.76

Table 6c
Grade 10 Mathematics Average Item Scores by Item Block
(Excludes Total CR Scores of 0)
(n=507)

Examination Reading 1

Item #	Pt. Value	Scoring Modality					
		1		2		3	
		Mean	SD	Mean	SD	Mean	SD
9	4	1.07	0.91	1.11	0.89	1.11	0.88
10	2	0.41	0.71	0.37	0.70	0.38	0.72
11	2	1.97	1.60	1.97	1.53	1.99	1.54
19	2	0.47	0.78	0.47	0.78	0.48	0.78
20	2	0.24	0.47	0.22	0.47	0.23	0.46
21	2	0.27	0.67	0.28	0.68	0.29	0.68
40	2	**0.47**	0.77	0.35	0.72	0.41	0.72
41	2	0.41	0.75	0.41	0.75	0.41	0.76
48	2	0.36	0.66	0.38	0.63	0.38	0.66
49	2	0.42	0.80	0.42	0.79	0.41	0.79
50	4	0.60	1.26	0.60	1.26	0.60	1.28

Examination Reading 2

Item #	Pt. Value	Scoring Modality					
		1		2		3	
		Mean	SD	Mean	SD	Mean	SD
9	4	1.10	0.89	1.08	0.89	1.09	0.89
10	2	**0.46**	0.75	0.35	0.65	0.38	0.70
11	2	2.05	1.59	1.96	1.50	2.02	1.53
19	2	0.48	0.78	0.47	0.78	0.46	0.78
20	2	0.24	0.48	0.21	0.44	0.22	0.45
21	2	0.28	0.69	0.28	0.68	0.27	0.67
40	2	**0.48**	0.78	0.36	0.72	0.42	0.75
41	2	0.42	0.76	0.41	0.76	0.41	0.75
48	2	0.36	0.67	0.39	0.65	0.37	0.65
49	2	0.42	0.80	0.42	0.79	0.40	0.77
50	4	0.59	1.26	0.59	1.24	0.58	1.25

Bolded values indicate differences in means SM1-SM2 or SM3-SM2 (2(SE) of SM2 mean
Boided italicized values indicate differences in means SM1-SM2 or SM3-SM2 -[-2(SE)] of SM2 mean

BEST COPY AVAILABLE

Table 7a
Reading Generalizability and D Studies for Scoring Modalities 1 and 2
"Multiple-Examination-Reading" Subsamples
(Excludes Total CR Scores of 0)

Grade 4

Scoring Modality 1		
Source of Variation	Est. Var. Component	% Total Variance
Person:rater	0.141	20.6
Rater	0.000*	0.0
Item	0.123	18.0
Item*rater	0.003	0.4
Residual	0.418	61.0
Tot. % Var.		100.0

n_r	1	2
n_i =	12	12
Rel. error var.	0.035	0.035
G. Coef.	**0.801**	**0.801**
Abs. error var.	0.045	0.045
Index of Dep.(Φ)	**0.757**	**0.757**

Scoring Modality 2		
Source of Variation	Est. Var. Component	% Total Variance
Person	0.137	20.6
Rater:item	0.000*	0.0
Item	0.113	17.0
Item*person	0.277	41.8
Residual	0.137	20.6
Tot. % Var.		100.0

n_r	2	4
n_i =	12	12
Rel. error var.	0.023	0.026
G. Coef.	**0.826**	**0.840**
Abs. error var.	0.038	0.035
Index of Dep.(Φ)	**0.781**	**0.794**

Grade 8

Scoring Modality 1		
Source of Variation	Est. Var. Component	% Total Variance
Person:rater	0.201	22.4
Rater	0.000*	0.0
Item	0.181	20.2
Item*rater	0.013	1.4
Residual	0.500	55.9
Tot. % Var.		100.0

n_r	1	2
n_i =	12	12
Rel. error var.	0.043	0.043
G. Coef.	**0.825**	**0.825**
Abs. error var.	0.058	0.058
Index of Dep.(Φ)	**0.776**	**0.776**

Scoring Modality 2		
Source of Variation	Est. Var. Component	% Total Variance
Person	0.187	21.5
Rater:item	0.000*	0.0
Item	0.178	20.4
Item*person	0.367	42.1
Residual	0.140	16.0
Tot. % Var.		100.0

n_r	2	4
n_i =	12	12
Rel. error var.	0.036	0.033
G. Coef.	**0.837**	**0.848**
Abs. error var.	0.051	0.048
Index of Dep.(Φ)	**0.785**	**0.795**

Grade 10

Scoring Modality 1		
Source of Varation	Est. Var. Component	% Total Variance
Person:rater	0.134	18.4
Rater	0.012	1.7
Item	0.140	19.2
Item*rater	0.003	0.5
Residual	0.440	60.2
Tot. % Var.		100.0

n_r	1	2
n_i =	11	11
Rel. error var.	0.040	0.040
G. Coef.	**0.784**	**0.784**
Abs. error var.	0.053	0.053
Index of Dep.(Φ)	**0.734**	**0.734**

Scoring Modality 2		
Source of Variation	Est. Var. Component	% Total Variance
Person	0.110	16.2
Rater:item	0.000*	0.0
Item	0.149	21.9
Item*person	0.311	45.8
Residual	0.110	16.1
Tot. % Var.		100.0

n_r	2	4
n_i =	11	11
Rel. error var.	0.033	0.031
G. Coef.	**0.768**	**0.782**
Abs. error var.	0.047	0.044
Index of Dep.(Φ)	**0.702**	**0.713**

* Negative variance component set to 0

BEST COPY AVAILABLE

Table 7b
Mathematics Generalizability and D Studies for Scoring Modalities 1 and 2
"Multiple-Examination-Reading" Subsamples
(Excludes Total CR Scores of 0)

Grade 5

Scoring Modality 1		
Source of Variation	Est. Var. Component	% Total Variance
Person:rater	0.170	17.5
Rater	0.000*	0.0
Item	0.232	23.8
Item*rater	0.003	0.3
Residual	0.569	58.4
Tot. % Var.		100.0

n_r	1	2
n_i =	11	11
Rel. error var.	0.052	0.052
G. Coef.	**0.766**	**0.766**
Abs. error var.	0.073	0.073
Index of Dep.(Φ)	**0.699**	**0.699**

Scoring Modality 2		
Source of Variation	Est. Var. Component	% Total Variance
Person	0.158	18.4
Rater:item	0.021	2.4
Item	0.147	17.0
Item*person	0.469	54.3
Residual	0.068	7.9
Tot. % Var.		100.0

n_r	2	4
n_i =	11	11
Rel. error var.	0.046	0.044
G. Coef.	**0.776**	**0.782**
Abs. error var.	0.060	0.045
Index of Dep.(Φ)	**0.725**	**0.731**

Grade 8

Scoring Modality 1		
Source of Variation	Est. Var. Component	% Total Variance
Person:rater	0.138	18.5
Rater	0.000*	0.0
Item	0.239	32.0
Item*rater	0.008	1.1
Residual	0.362	48.4
Tot. % Var.		100.0

n_r	1	2
n_i =	11	11
Rel. error var.	0.034	0.034
G. Coef.	**0.804**	**0.804**
Abs. error var.	0.055	0.055
Index of Dep.(Φ)	**0.714**	**0.714**

Scoring Modality 2		
Source of Variation	Est. Var. Component	% Total Variance
Person	0.129	17.2
Rater:item	0.000*	0.0
Item	0.235	31.3
Item*person	0.331	44.2
Residual	0.054	7.3
Tot. % Var.		100.0

n_r	2	4
n_i =	11	11
Rel. error var.	0.033	0.031
G. Coef.	**0.799**	**0.805**
Abs. error var.	0.054	0.053
Index of Dep.(Φ)	**0.705**	**0.710**

Grade 10

Scoring Modality 1		
Source of Variation	Est. Var. Component	% Total Variance
Person:rater	0.193	17.8
Rater	0.001	0.1
Item	0.264	24.4
Item*rater	0.000	0.0
Residual	0.626	57.7
Tot. % Var.		100.0

n_r	1	2
n_i =	11	11
Rel. error var.	0.057	0.057
G. Coef.	**0.773**	**0.773**
Abs. error var.	0.081	0.081
Index of Dep.(Φ)	**0.706**	**0.706**

Scoring Modality 2		
Source of Variation	Est. Var. Component	% Total Variance
Person	0.224	19.5
Rater:item	0.015	1.3
Item	0.248	21.7
Item*person	0.516	44.9
Residual	0.145	12.6
Tot. % Var.		100.0

n_r	2	4
n_i =	11	11
Rel. error var.	0.053	0.050
G. Coef.	**0.807**	**0.817**
Abs. error var.	0.077	0.073
Index of Dep.(Φ)	**0.745**	**0.754**

* Negative variance component set to 0

BEST COPY AVAILABLE